

MiniSearchEngine—迷你搜索引擎

高级数据结构与算法分析 Project1 实验报告

第 19 小组：邵轩溥，王子文，何宇凡

指导教师：卜佳俊

完成时间：2024 年 3 月 21 日

一、问题与背景介绍

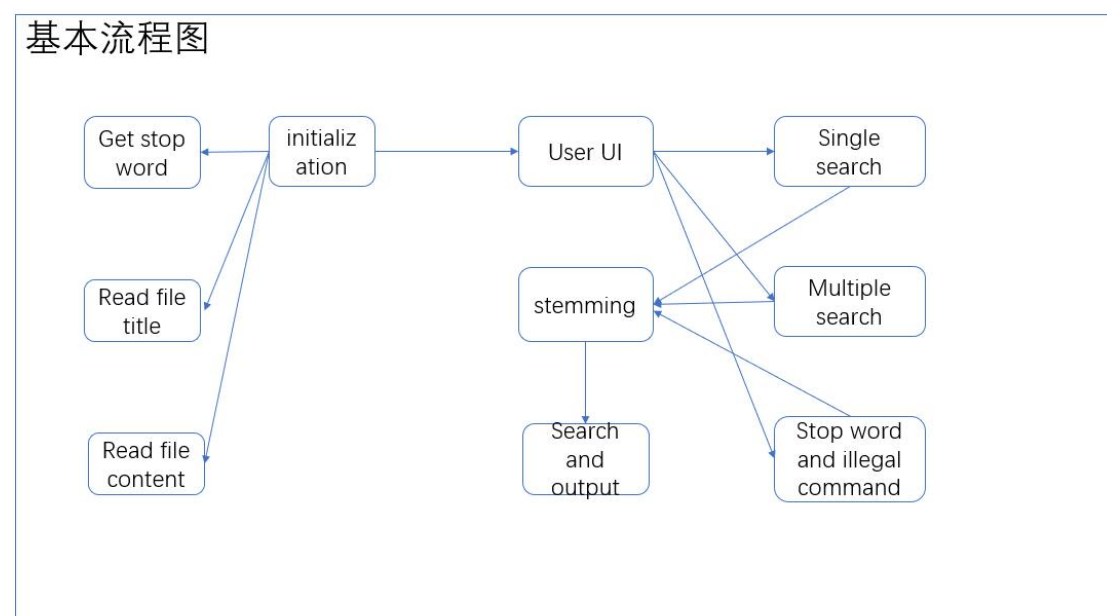
随着信息技术的不断发展，文本信息的存储和检索成为了人们日常生活和工作中的重要需求。在信息检索领域，倒排索引是一种常用且高效的数据结构，用于快速定位包含特定词项的文档集合。倒排索引是实现“单词-文档”矩阵的一种具体存储形式，通过倒排索引，可以根据单词快速获取包含这个单词的文档列表，为文本检索提供了强大的支持。

搜索引擎的实现就是通过倒排索引。当用户发起查询时，搜索引擎扫描索引库中的所有文档，找出所有包含关键词的文档，这样依次从文档中去查找是否含有关键词的方法叫做正向索引。对于数量极大的文档，正向索引的效率十分低下无法满足用户需求，因此需要使用倒排索引。

在本次实验中，我们将实现一个基于“莎士比亚全集”的迷你搜索引擎。对于给定的一个查询词，搜索引擎可以给出它出现在哪个文档中并给出它出现的次数；对于给定的一组查询词，搜索引擎则输出所有查询词都出现的文档。

二、算法介绍

2.1 基本流程图



2.2 伪代码

```

int main():
    Call function getstopword() // Get the stopword list
    Call function readfiletitle() // Read document titles
    Call function readfilecontent() // Read document content and build inverted index

    Label home: // Return here after each non-quit operation
        Print message: "Here we read all the texts you provided, Here are ", the number of distinct
words in the inverted index, " distinct words"
        Print message: "The size of stop word list is ", the size of the stopword list
        Print message: "Press S to search single word, press M to search multiple words, press Q to exit:"

    Read user input and store it in variable t
    If t is 'S' or 's': // Single word search
        Read user input and store it in variable word
        Perform stemming on word
        If word is a stopword:
            Print message: "You have entered a stop word"
            Go to label home
        If word exists in the inverted index:
            Get the inner map corresponding to word and iterate through it:
                Print the number of times word appears in each file along with the filename
        Else:
            Print message: "Word not found in inverted index."
            Print message: "Press any key to back to the search"
            Read user input and store it in variable w
            Go to label home
    Else, if t is 'q' or 'Q': // Quit the program
        Exit the program
    Else, if t is 'm' or 'M': // Multiple word search
        Ignore newline characters in user input
        Read user input and store it in variable s
        Split s by spaces and store in string array words
        Perform stemming on each word in words
        Initialize an array appearance to store the number of times each file appears, all set to 0
        Initialize variable i to 0
        For each word in words:
            Get the inner map corresponding to word and iterate through it:
                Update the appearance count for the corresponding file
            Increment the value of i
        Initialize variable isresult to 0
        For each file k in filename:
            If appearance[k] equals i: // Check if all words appear in the same file
                Print message: "Those words all appears in ", current filename
                Set isresult to 1
        If isresult is 0:
            Print message: "No results, back to the homepage"
            Go to label home
        Else:
            Print message: "Illegal Command! "
            Go to label home
    Call system command "Pause"

```

2.3 倒排索引

倒排索引用于存储每个单词出现的文章索引。其中，主键是单词（字符串类型），键值是一个列表，保存了单词出现过的所有文章的独立标签，以便于进行短语搜索，可以保存“文章标签 - 文章中单词位置”的对应关系。

```
std::map<std::string,std::map<int,int>>>invertedindex;
```

2.4 构建停用词列表

该算法初始化一组停用词，用于处理文档和查询中的术语。停用词是应从索引中排除以提高搜索效率的常见单词，我们是直接从网上下载。

```
std::set<std::string> stopwordslist;
```

2.5 储存所有文件名

```
std::vector<std::string> filename;
```

2.6 储存所有词

储存出现在文件中的所有词。

```
std::set<std::string> totalwords;
```

2.7 词干程序

这段代码调用了一个名为 Porter2Stemmer 的类中的 trim 和 stem 方法来对字符串进行处理。trim 方法用于去除字符串两端的空白字符，而 stem 方法则用于对字符串进行词干提取处理，即将单词转换为其基本形式。最后，处理后的字符串被返回。对于词干提取器函数，请打开 porter2_stemmer.cpp 和它的头文件以获取更多细节（感谢这个函数，它确实为我们节省了一些时间）。

```
std::string stemmer(std::string s){
    Porter2Stemmer::trim(s);
    Porter2Stemmer::stem(s);
    return s;
}
```

2.8 其他伪代码

Function getstopword():

Set file to "stopword.txt" // Specify the file name

Open the file for reading and store the file stream in 'in'

If the file is successfully opened:

Initialize an empty string 'word'

```
    While there are words to be read from the file:
        Read a word from the file and store it in 'word'
        Insert the word into the stopwordlist set
    Close the file
    Print message: "Stop words have been loaded."
Else:
    Print error message: "Unable to open the file: " followed by the file name
```

Function readfiletitle():

```
Set file to "titles.txt" // Specify the file name
Open the file for reading and store the file stream in 'in'
If the file is successfully opened:
    Initialize an empty string 'word'
    While there are words to be read from the file:
        Read a word from the file and store it in 'word'
        Add 'word' to the end of the filename vector
    Close the file
    Print message: "Titles have been loaded."
Else:
    Print error message: "Unable to open the file: " followed by the file name
```

Function isstopword(s):

```
If stopwordlist is empty:
    Call function getstopword() // Ensure stopwordlist is populated
If s is not found in stopwordlist:
    Return true // s is a stopword
Else:
    Return false // s is not a stopword
```

Function readfilecontent():

```
For each file in filename:
    Set path to "FullShakespeare/" concatenated with the current filename
    Open the file specified by 'path' for reading and store the file stream in 'in'
    If the file is successfully opened:
        Initialize an empty string 'word'
        While there are words to be read from the file:
            Read a word from the file and store it in 'word'
            Apply stemming to 'word'
            If 'word' is not a stopword:
                Continue to the next iteration of the loop
            If 'word' is not found in totalwords:
                Add 'word' to the totalwords set
            If 'word' is not found in invertedindex:
                Add a new entry to invertedindex with 'word' as key and an empty
```

map as value

Increment the count of 'word' in the current file in invertedindex

Else:

Print error message: "Unable to open the file: " followed by the current filename

三、测试结果

初始化完成。

```
Stop words have been loaded.
Titles have been loaded.
Here we read all the texts you provided, Here are 18309 distinct words
The size of stop word list is 851
Press S to search single word, press M to search multiple words, press Q to exit:
```

单单词查询测试样例：

这里测试一些常用的、不常用的、未出现的、比较有代表性的单词以及停用词。

常用的单词 book

```
Press S to search single word, press M to search multiple words, press Q to exit:s
book
In Book: 1henryiv.txt,It appears: 4times
In Book: 1henryvi.txt,It appears: 3times
In Book: 1kinghenryiv.txt,It appears: 4times
In Book: 1kinghenryvi.txt,It appears: 3times
In Book: 2henryiv.txt,It appears: 8times
In Book: 2henryvi.txt,It appears: 5times
In Book: 2kinghenryiv.txt,It appears: 8times
In Book: 2kinghenryvi.txt,It appears: 5times
In Book: 3henryvi.txt,It appears: 1times
In Book: 3kinghenryvi.txt,It appears: 1times
In Book: asyoulikeit.txt,It appears: 4times
In Book: cleopatra.txt,It appears: 1times
In Book: coriolanus.txt,It appears: 2times
In Book: cymbeline.txt,It appears: 3times
In Book: hamlet.txt,It appears: 3times
In Book: henryv.txt,It appears: 1times
In Book: henryviii.txt,It appears: 1times
In Book: john.txt,It appears: 3times
In Book: julius_caesar.txt,It appears: 2times
In Book: lear.txt,It appears: 1times
In Book: lll.txt,It appears: 9times
In Book: macbeth.txt,It appears: 1times
In Book: measure.txt,It appears: 1times
In Book: merchant.txt,It appears: 2times
In Book: merry_wives.txt,It appears: 7times
In Book: midsummer.txt,It appears: 1times
In Book: much_ado.txt,It appears: 5times
In Book: othello.txt,It appears: 2times
```

```

In Book: pericles.txt,It appears: 2times
In Book: richardii.txt,It appears: 3times
In Book: richardiii.txt,It appears: 4times
In Book: romeoandjuliet.txt,It appears: 8times
In Book: romeo_juliet.txt,It appears: 8times
In Book: taming_shrew.txt,It appears: 12times
In Book: tempest.txt,It appears: 8times
In Book: timon.txt,It appears: 2times
In Book: titus.txt,It appears: 5times
In Book: troilus_cressida.txt,It appears: 2times
In Book: twelfth_night.txt,It appears: 3times
In Book: two_gentlemen.txt,It appears: 1times
In Book: winters_tale.txt,It appears: 1times
Press any key to back to the search

```

使用 windows 的 findstr 命令查找进行验证

```

PS E:\5SpringAndSummerCourse\ads\project1\MiniSearchEngine\FullShakespeare> findstr /s /r "book" *.*
1henryiv.txt:And now I will unclasp a secret book,
1henryiv.txt:O Lord, sir, I'll be sworn upon all the books in
1henryiv.txt:By that time will our book, I think, be drawn
1henryiv.txt:By this our book is drawn; we'll but seal,
1henryvi.txt:Unless my study and my books be false,
1henryvi.txt:I'll note you in my book of memory,
1henryvi.txt:And fitter is my study and my books
1kinghenryiv.txt:And now I will unclasp a secret book,
1kinghenryiv.txt:O Lord, sir, I'll be sworn upon all the books in
1kinghenryiv.txt:By that time will our book, I think, be drawn
1kinghenryiv.txt:By this our book is drawn; we'll but seal,
1kinghenryvi.txt:Unless my study and my books be false,
1kinghenryvi.txt:I'll note you in my book of memory,
1kinghenryvi.txt:And fitter is my study and my books
2henryiv.txt:book-oath: deny it, if thou canst.
2henryiv.txt:book as thou and Falstaff for obduracy and
2henryiv.txt:He was the mark and glass, copy and book,
2henryiv.txt:lipping to his master's old tables, his note-book,
2henryiv.txt:O God! that one might read the book of fate,
2henryiv.txt:Would shut the book, and sit him down and die.
2henryiv.txt:a number of shadows to fill up the muster-book.
2henryiv.txt:Turning your books to graves, your ink to blood,
2henryiv.txt:That you should seal this lawless bloody book
2henryiv.txt:How deep you were within the books of God?
2henryiv.txt:I beseech your grace, let it be booked with the
2henryvi.txt:Blotting your names from books of memory,
2henryvi.txt:Whose bookish rule hath pull'd fair England down.

```

(略去中间部分)

```

tempest.txt:Knowing I loved my books, he furnish'd me
tempest.txt:Here, kiss the book. Though thou canst swim like a
tempest.txt:Come, swear to that; kiss the book: I will furnish
tempest.txt:At nothing can be more. I'll to my book,
tempest.txt:Having first seized his books, or with a log
tempest.txt:First to possess his books; for without them
tempest.txt:As rootedly as I. Burn but his books.
tempest.txt:I'll drown my book.
timon.txt:A picture, sir. When comes your book forth?
timon.txt:Pays interest for 't; his land's put to their books.
titus.txt:Enter young LUCIUS, and LAVINIA running after him, and the boy flies from her, with books under his arm. Then
enter TITUS and MARCUS
titus.txt:Which made me down to throw my books, and fly--
titus.txt:LAVINIA turns over with her stumps the books which LUCIUS has let fall
titus.txt:Some book there is that she desires to see.
titus.txt:Lucius, what book is that she tosseth so?
troilus_cressida.txt:thou learn a prayer without book. Thou canst strike,
troilus_cressida.txt:O, like a book of sport thou'lt read me o'er;
twelfth_night.txt:word for word without book, and hath all the good
twelfth_night.txt:To thee the book even of my secret soul:
twelfth_night.txt:that cons state without book and utters it by great
two_gentlemen.txt:And on a love-book pray for my success?
two_gentlemen.txt:Upon some book I love I'll pray for thee.
winters_tale.txt:though I am not bookish, yet I can read
winters_tale.txt:put in the book of virtue!
winters_tale.txt:ribbon, glass, pomander, brooch, table-book, ballad,

```

与我们的结果吻合。

不常用的：apple


```

Press S to search single word, press M to search multiple words, press Q to exit:s
apple
In Book: henryv.txt,It appears: 1times
In Book: henryviii.txt,It appears: 1times
In Book: lear.txt,It appears: 1times
In Book: lll.txt,It appears: 1times
In Book: merchant.txt,It appears: 1times
In Book: midsummer.txt,It appears: 1times
In Book: taming_shrew.txt,It appears: 2times
In Book: tempest.txt,It appears: 1times
In Book: twelfth_night.txt,It appears: 2times
Press any key to back to the search

```

同样进行验证

```

PS E:\5SpringAndSummerCourse\ads\project1\MiniSearchEngine\FullShakespeare> findstr /s /r "apple" *.*
1henryiv.txt:And let them grapple: O, the blood more stirs
1henryiv.txt:gown; I am withered like an old apple-john. Well,
1kinghenryiv.txt:And let them grapple: O, the blood more stirs
1kinghenryiv.txt:gown; I am withered like an old apple-john. Well,
2henryiv.txt:What the devil hast thou brought there? apple-johns?
2henryiv.txt:thou knowest Sir John cannot endure an apple-john.
2henryiv.txt:of apple-johns before him, and told him there were
2henryvi.txt:To grapple with the house of Lancaster;
2kinghenryiv.txt:What the devil hast thou brought there? apple-johns?
2kinghenryiv.txt:thou knowest Sir John cannot endure an apple-john.
2kinghenryiv.txt:of apple-johns before him, and told him there were
2kinghenryvi.txt:To grapple with the house of Lancaster;
asyoulikeit.txt:And yet it irks me the poor dappled fools,
hamlet.txt:Grapple them to thy soul with hoops of steel;
hamlet.txt:a compelled valour, and in the grapple I boarded
henryv.txt:Grapple your minds to sternage of this navy,
henryv.txt:rotten apples! You may as well say, that's a
henryviii.txt:and fight for bitten apples; that no audience, but
john.txt:And grapple with him ere he comes so nigh.
john.txt:And grapple thee unto a pagan shore;
lear.txt:apple, yet I can tell what I can tell.
lll.txt:I was as willing to grapple as he was to board.
lll.txt:And laugh upon the apple of her eye?
macbeth.txt:Grapples you to the heart and love of us,
merchant.txt:A goodly apple rotten at the heart:
midsummer.txt:Sink in apple of his eye.
much_ado.txt:Dapples the drowsy east with spots of grey.
taming_shrew.txt:apples. But come; since this bar in law makes us
taming_shrew.txt:[Aside] As much as an apple doth an oyster,

```

结果吻合。

未出现的词 toufu

```

Press S to search single word, press M to search multiple words, press Q to exit:s
toufu
Word not found in inverted index.
Press any key to back to the search

```

验证结果

```

PS E:\5SpringAndSummerCourse\ads\project1\MiniSearchEngine\FullShakespeare> findstr /s /r "toufu" *.*
PS E:\5SpringAndSummerCourse\ads\project1\MiniSearchEngine\FullShakespeare> |

```

代表性的词 hamlet

```

Press S to search single word, press M to search multiple words, press Q to exit:s
hamlet
In Book: hamlet.txt,It appears: 473times
Press any key to back to the search

```

只在 hamlet.txt 中出现很多次，与预测是符合的。

停用词 never

```

Press S to search single word, press M to search multiple words, press Q to exit:s
never
You have entered a stop word
Here we read all the texts you provided, Here are 18309 distinct words
The size of stop word list is 851
Press S to search single word, press M to search multiple words, press Q to exit:|

```

多单词查询测试样例：

read book

```
Press S to search single word, press M to search multiple words, press Q to exit:m
read book
Those words all appears in 1henryiv.txt
Those words all appears in 1henryvi.txt
Those words all appears in 1kinghenryiv.txt
Those words all appears in 1kinghenryvi.txt
Those words all appears in 2henryiv.txt
Those words all appears in 2henryvi.txt
Those words all appears in 2kinghenryiv.txt
Those words all appears in 2kinghenryvi.txt
Those words all appears in 3henryvi.txt
Those words all appears in 3kinghenryvi.txt
Those words all appears in asyoulikeit.txt
Those words all appears in cleopatra.txt
Those words all appears in coriolanus.txt
Those words all appears in cymbeline.txt
Those words all appears in hamlet.txt
Those words all appears in henryv.txt
Those words all appears in henryviii.txt
Those words all appears in john.txt
Those words all appears in julius_caesar.txt
Those words all appears in lear.txt
Those words all appears in lll.txt
Those words all appears in macbeth.txt
Those words all appears in measure.txt
Those words all appears in merchant.txt
Those words all appears in merry_wives.txt
```

desire move

```
Press S to search single word, press M to search multiple words, press Q to exit:m
desire move
No results, back to the homepage
Here we read all the texts you provided, Here are 18312 distinct words
The size of stop word list is 851
```

King sad command

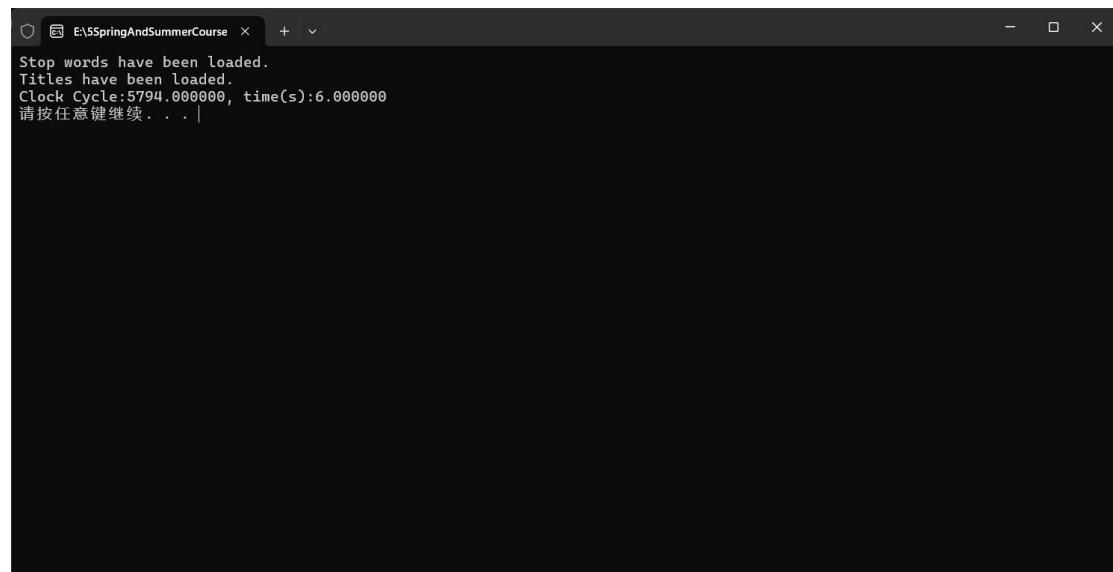
```
Press S to search single word, press M to search multiple words, press Q to exit:m
king sad command
Those words all appears in 1henryiv.txt
Those words all appears in 1henryvi.txt
Those words all appears in 1kinghenryiv.txt
Those words all appears in 1kinghenryvi.txt
Those words all appears in 2henryiv.txt
Those words all appears in 2henryvi.txt
Those words all appears in 2kinghenryiv.txt
Those words all appears in 2kinghenryvi.txt
Those words all appears in 3henryvi.txt
Those words all appears in 3kinghenryvi.txt
Those words all appears in allswell.txt
Those words all appears in allswellthatendswell.txt
Those words all appears in asyoulikeit.txt
Those words all appears in cleopatra.txt
Those words all appears in coriolanus.txt
Those words all appears in cymbeline.txt
Those words all appears in hamlet.txt
Those words all appears in henryv.txt
Those words all appears in henryviii.txt
Those words all appears in john.txt
Those words all appears in julius_caesar.txt
Those words all appears in lear.txt
Those words all appears in lll.txt
Those words all appears in macbeth.txt
Those words all appears in measure.txt
Those words all appears in merchant.txt
Those words all appears in merry_wives.txt
```

```
Those words all appears in much_ado.txt
Those words all appears in othello.txt
Those words all appears in pericles.txt
Those words all appears in richardii.txt
Those words all appears in richardiii.txt
Those words all appears in taming_shrew.txt
Those words all appears in tempest.txt
Those words all appears in titus.txt
Those words all appears in troilus_cressida.txt
Those words all appears in twelfth_night.txt
Those words all appears in two_gentlemen.txt
Those words all appears in winters_tale.txt
Here we read all the texts you provided, Here are 18312 distinct words
The size of stop word list is 851
```

速度测试：

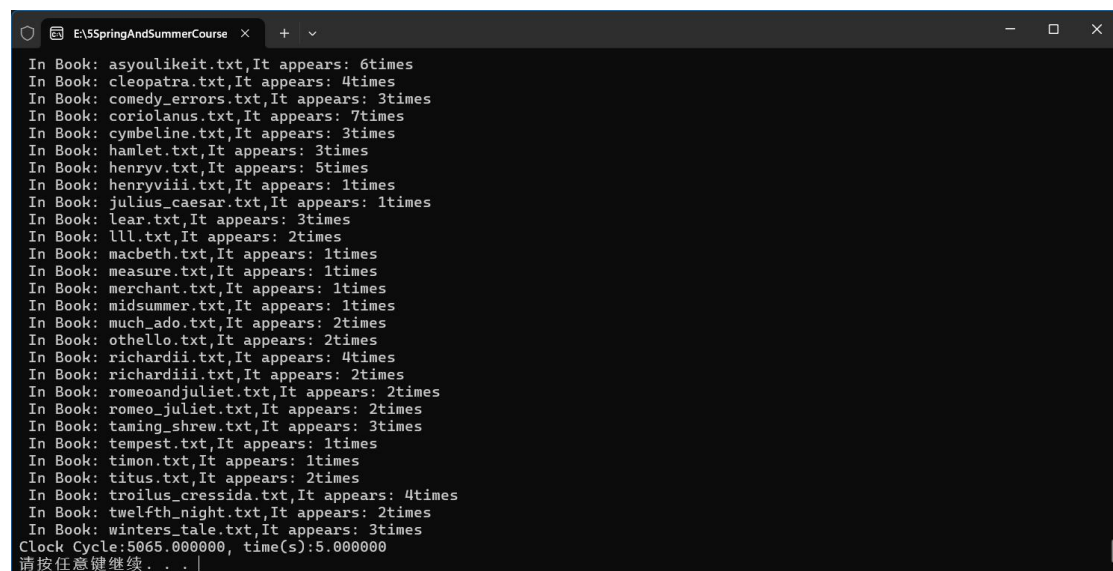
封装代码后进行速度测试

初始化用时 将 stopwords, title, content 都读入花费 5794 个时钟周期，6s 完成。



```
E:\SSpringAndSummerCourse x + v
Stop words have been loaded.
Titles have been loaded.
Clock Cycle:5794.000000, time(s):6.000000
请按任意键继续. . .
```

(单词) 1000 次查询花费 5065 个时钟周期，5s 钟完成。



```
E:\SSpringAndSummerCourse x + v
In Book: asyoulikeit.txt,It appears: 6times
In Book: cleopatra.txt,It appears: 4times
In Book: comedy_errors.txt,It appears: 3times
In Book: coriolanus.txt,It appears: 7times
In Book: cymbeline.txt,It appears: 3times
In Book: hamlet.txt,It appears: 3times
In Book: henryv.txt,It appears: 5times
In Book: henryviii.txt,It appears: 1times
In Book: julius_caesar.txt,It appears: 1times
In Book: lear.txt,It appears: 3times
In Book: lll.txt,It appears: 2times
In Book: macbeth.txt,It appears: 1times
In Book: measure.txt,It appears: 1times
In Book: merchant.txt,It appears: 1times
In Book: midsummer.txt,It appears: 1times
In Book: much_ado.txt,It appears: 2times
In Book: othello.txt,It appears: 2times
In Book: richardii.txt,It appears: 4times
In Book: richardiii.txt,It appears: 2times
In Book: romeoandjuliet.txt,It appears: 2times
In Book: romeo_juliet.txt,It appears: 2times
In Book: taming_shrew.txt,It appears: 3times
In Book: tempest.txt,It appears: 1times
In Book: timon.txt,It appears: 1times
In Book: titus.txt,It appears: 2times
In Book: troilus_cressida.txt,It appears: 4times
In Book: twelfth_night.txt,It appears: 2times
In Book: winters_tale.txt,It appears: 3times
Clock Cycle:5065.000000, time(s):5.000000
请按任意键继续. . .
```

(多单词) 1000 条查询项，包含 2-4 个单词

```
E:\SpringAndSummerCourse x + v
wooden word work worker
No results, back to the homepage
working works
No results, back to the homepage
workshop world worried
No results, back to the homepage
worry worth would wound
No results, back to the homepage
wrap write
Those words all appears in 3henryvi.txt
Those words all appears in 3kinghenryvi.txt
Those words all appears in asyoulikeit.txt
Those words all appears in hamlet.txt
Those words all appears in lear.txt
Those words all appears in midsummer.txt
Those words all appears in titus.txt
writer writing wrong
No results, back to the homepage
yard yeah year yell
No results, back to the homepage
yellow yes
No results, back to the homepage
yesterday yet yield
No results, back to the homepage
you young your yours
No results, back to the homepage
yourself youth
No results, back to the homepage
Clock Cycle:325.000000, time(s):0.000000
请按任意键继续. . . |
```

四、复杂度分析

The map and set of C++ STL are implemented by RBT.

4.1 时间复杂度

void getstopword(): $O(n)$

文件读取: $O(n)$,

插入操作: $O(\log n)$

void loadFileNames(): $O(n)$

文件读取: $O(n)$

插入操作: $O(\log n)$

bool isstopword(std::string s): $O(n)$

void readfilecontent(): $O(m*n*\log n)$;

打开文件: $O(1)$,

读取文件内容: $O(m)$ 。

插入和查找操作: $O(\log n)$

4.2 空间复杂度

void getstopword(): $O(1)$

简单地从文件中读取内容, 不需要额外的空间。

void loadFilenames(): $O(1)$

简单地从文件中读取内容，不需要额外的空间。

bool isstopword(std::string s) : $O(1)$

不需要额外空间。

void readfilecontent(): $O(n)$

空间复杂度主要由存储所有单词的集合和映射所需的空间决定。

五、附录

一些重要源码：

初始化（读取停用词，标题，文本内容）

```
void getstopword() {
    std::string file = "stopword.txt"; // open the file
    std::ifstream in(file);

    if (in) {
        std::string word;
        while (in >> word) {
            stopwordlist.insert(word); // insert it one by one
        }
        in.close();
        std::cout << "Stop words have been loaded." << std::endl;
    } else {
        std::cerr << "Unable to open the file: " << file << std::endl; // error
    }
}

void readfiletitle(){
    std::string file = "titles.txt";
    std::ifstream in(file);
    if (in) {
        std::string word;
        while (in >> word) {
            filename.push_back(word);
        }
        in.close();
        std::cout << "Titles have been loaded." << std::endl;
    } else {
        std::cerr << "Unable to open the file: " << file << std::endl;
    }
}
```

```

    }
}
void readfilecontent(){
    for(int i=0;i<filename.size();i++){
        std::string path="FullShakespeare//";
        std::ifstream in(path+filename[i]);
        if(in){
            std::string word;
            while(in>>word) {
                word=stemmer(word);//stem the word first
                if(isstopword(word)==false){
                    continue;
                }
                if(totalwords.find(word)==totalwords.end()){
                    totalwords.insert(word);
                }
                if (invertedindex.find(word)==invertedindex.end()) {
                    // 如果外层键不存在，则添加一个新的内层 map
                    invertedindex[word]=std::map<int, int>();
                }
                invertedindex[word][i]++;
            }
        }else{
            std::cerr << "Unable to open the file: " << filename[i] <<
std::endl;
        }
    }
}

```

主函数:

```

int main(){
    getstopword();
    readfiletitle();
    readfilecontent();
    home: //Every time a non-quit operation is over,we will be back here.
    std::cout<<"Here we read all the texts you provided, Here are ";
    std::cout<<invertedindex.size()<<" distinct words"<<std::endl;
    std::cout<<"The size of stop word list is
"<<stopwordlist.size()<<std::endl;
    std::cout<<"Press S to search single word, press M to search multiple
words, press Q to exit:";
    char t;
    std::cin>>t;
    if(t=='S' || t=='s'){//single word search

```

```

        std::string word;
    std::cin>>word;
    word=stemmer(word);
    if(isstopword(word)==false){
        std::cout<<"You have entered a stop word"<<std::endl;
        goto home;
    }
    if (invertedindex.find(word)!=invertedindex.end()) {
        std::map<int, int>& inner_map = invertedindex[word];
        for (const auto& pair : inner_map) {
            std::cout << " In Book: " << filename[pair.first] << ",It appears:"
" << pair.second << "times" <<std::endl;
        }
    } else {
        std::cout << "Word not found in inverted index."<<std::endl;
    }

    std::cout << "Press any key to back to the search"<<std::endl;
    char w;
    std::cin>>w;
    goto home;
}else if(t=='q' || t=='Q'){//quit the program
    exit(-1);
}else if(t=='m' || t=='M'){//multiple words search
    std::cin.ignore();
    std::string s;
    std::vector<int> appearance(filename.size(), 0);
    int i=0;
    std::getline(std::cin,s);
    std::stringstream ss(s);
    std::vector<std::string>words;
    std::string token;
    while (ss >>token) {
        stemmer(token);//To handle input correstly, stemming is also
necessary.
        words.push_back(token);
    }
    while(i<words.size()){
        std::map<int, int>& inner_map = invertedindex[words[i]];
        for (const auto& pair : inner_map){
            appearance[pair.first]++;
        }
        i++;
    }
    int isresult=0;

```

```

        for(int k=0;k<filename.size();k++){
            if(appearance[k]==i){//This is to check whether all the words
appears in the same file.
                std::cout<<"Those words all appears in
"<<filename[k]<<std::endl;
                isresult=1;
            }
        }
        if(isresult==0){
            std::cout<<"No results, back to the homepage"<<std::endl;
        }
        goto home;
    }else{
        std::cout<<"Illegal Command! "<<std::endl;
        goto home;
    }
    system("Pause");
}

```

六、引用

Github 上开源的词干提取器 C++ Porter2_stemmer

停用词表

851 词

七、声明

We hereby declare that all the work done in this project titled "Roll Your Own Mini Search Engine" is of our independent effort as a group.

Signature: 邵轩溥, 王子文, 何宇凡